Evaluating comprehension of natural and synthetic conversational speech

Evaluating comprehension of natural and synthetic conversational speech

Mirjam Wester, Oliver Watts and Gustav Eje Henter

The Centre for Speech Technology Research, The University of Edinburgh, United Kingdom

mwester@inf.ed.ac.uk

Abstract

Current speech synthesis methods typically operate on isolated sentences and lack convincing prosody when generating longer segments of speech. Similarly, prevailing TTS evaluation paradigms, such as intelligibility (transcription word error rate) or MOS, only score sentences in isolation, even though overall comprehension arguably is more important for speech-based communication. In an effort to develop more ecologically-relevant evaluation techniques that go beyond isolated sentences, we investigated comprehension of natural and synthetic speech dialogues. Specifically, we tested listener comprehension on long segments of spontaneous and engaging conversational speech (three 10-minute radio interviews of comedians). Interviews were reproduced either as natural speech, synthesised from carefully prepared transcripts, or synthesised using durations from forced-alignment against the natural speech, all in a balanced design. Comprehension was measured using multiple choice questions. A significant difference was measured between the comprehension/retention of natural speech (74% correct responses) and synthetic speech with forced-aligned durations (61% correct responses). However, no significant difference was observed between natural and regular synthetic speech (70% correct responses). Effective evaluation of comprehension remains elusive.

Index Terms: evaluation, comprehension, conversational speech, statistical parametric speech synthesis

1. Introduction

The goal of this work is to develop more ecologically-valid evaluation techniques that go beyond isolated sentences and measure *comprehension* of synthetic speech compared to natural speech. Modern text-to-speech synthesis systems (Statistical Parametric Speech Synthesis (SPSS)) produce speech that is comparable to natural speech in terms of intelligibility [1, 2]. Intelligibility, in this context, is mostly measured by transcription of semantically unpredictable sentences (SUS). However, the suprasegmental factors (duration, intonation, accents, pauses, etc.) that play a large role in comprehension do not figure prominently in the decoding of SUS (meaningless sentences in isolation). Therefore, this type of intelligibility may not be a good indicator of the comprehensibility of synthetic speech.

Prosody (suprasegmental factors) is arguably the defining difference between natural and synthetic speech. The importance of prosody for natural speech comprehension has been extensively investigated (see [3] for an overview). At the same time, the lack of appropriate prosody in TTS has been shown to affect speech intelligibility, response latencies and comprehension [4, 5, 6]. The effect of suprasegmental aspects on the comprehension of speech is not measured effectively by either MOS or SUS. The evaluation proposed in this paper attempts to fill this void.

A further motivation for developing better evaluation methods is the type of data that is increasingly being exploited as a source of speech data for SPSS – specifically 'found data' such as audio books or broadcast media data [1, 7, 8, 9, 10]. This is used, for instance, to create more expressive synthesis [7, 8] or for multi-lingual lightly-supervised TTS [9]. However, the evaluation methods needed to measure the success of these approaches are still lacking.

The data considered here – Desert Island Discs interviews – have interesting and meaningful prosody, are engaging to listen to, and are full of information. The evaluation paradigm used to measure comprehension is multiple-choice questions. The next section discusses prior work in this field and how our study fits with this literature.

2. Prior work

Attempts at developing comprehension tests to evaluate speech synthesis have a long history. Comprehension or comprehensibility of synthetic speech has been measured in a variety of different ways ranging from sentence verification tasks [11, 12] via immediate recall [13, 5], word-monitoring [14] multiple-choice questions [15, 16, 17, 18, 6, 19] and a summarization task [20], to computing math sums [21].

Previous attempts (in the 80s and 90s) at measuring comprehension using post-perceptual measures often did not show significant differences in comprehension between synthetic (formant-based) and natural speech [15, 16, 18]. (See [22] for a comprehensive review of comprehension of synthetic speech produced by rule.) [11] suggests this is due to the type of tests used. By using post-perceptual measures and multiple-choice questions or recall measures, subjects are encouraged to exploit real-world knowledge to solve the task. [11] argues that more sensitive measures are needed that measure the online perception process and proposed using the sentence verification task (SVT). Online measures of perceptual processing show that there are perceptual difficulties in interpreting high quality synthetic speech [11, 12] which disappear by the time the entire comprehension process has run its course [23].

Online methods generally use sentence-level materials which have been carefully constructed, for instance, to control for predictability [11] or text difficulty [14]. Evaluation techniques that are suitable for found data need to be able to evaluate longer stretches of speech, e.g., dialogues or stories. Therefore, online methods do not appear to be appropriate for e.g., Desert Island Discs data.

More recent studies investigating comprehension of synthetic speech (unit-selection and SPSS) revisit the use of multiple-choice questions [6, 19]. In [6] the comprehension of natural and unit-selection synthetic speech is compared using mainly multiple-choice questions. They found a significant difference in comprehension with synthetic speech scoring lower,

whereas intelligibility, measured using SUS, was equal between the two types of speech. Modifying the synthetic speech by inserting pauses between intonational phrases, thereby approximating the pauses in natural speech, removed the difference in comprehension between natural and synthetic speech. Chang [19] explored the relationship between intelligibility and comprehension of unit-selection and HMM-based synthesis. However, despite differences in intelligibility, no significant differences in comprehension were found when listening to either natural or synthetic speech.

The results in [19] and to a certain extent [6] fit an interpretation that post-perceptual tests are possibly not sufficiently sensitive to measure differences in comprehension. Nevertheless, we feel the post-perceptual approaches to measuring comprehension are not yet exhausted. There are a number of factors that we suspect may be affecting the sensitivity of post-perceptual tests. Firstly, the length of the speech material might be too short, i.e, 12–31 words [16, 17], or under 2 minutes [18]. Secondly, the content might not be prosodically interesting enough, i.e., rather boring sentences selected from news articles [19] or reading comprehension tests [18]. Finally, the questions asked may not be of the right type, i.e., higher-level questions can be answered even with poor intelligibility scores [19].

Our experiment is different to previous work in terms of the type and duration of the speech material and the type of questions. The Desert Island Discs material that we use is:

- prosodically rich; it comprises interesting and engaging interviews with comedians,
- 10 minutes long for each interview,
- tested using multiple choice questions which are surface structure or low-proposition questions, i.e., the participants are required to recall exact wording or detailed information about the speech content, thus not additionally relying on real-world knowledge.

Our expectation is that controlling these factors should increase the sensitivity of post-perceptual testing for measuring differences in comprehension between synthetic and natural speech.

3. Method

3.1. Evaluation data

"Desert Island Discs" is a BBC Radio 4 programme [24]. The programme is in the format of an interview, in which a guest is invited by Kirsty Young (the host) to choose the eight records they would take with them to a desert island. Three episodes were selected for the evaluation. The guests in the three episodes are all British comedians: David Walliams, Steve Coogan and Victoria Wood. From each of the interviews, we extracted 10 minutes of speech. Twenty multiple-choice questions were created for each excerpt. The questions all refer to the exact wording or detailed information about the speech content. Table 1 shows two questions asked about the David Walliams episode.

3.2. Experimental set-up

Our experiment consisted of three interviews (David Walliams (DW), Steve Coogan (SC) and Victoria Wood (VW)) reproduced using three speech types (natural (N), synthetic (S) and synthetic-modified (M)). The two synthetic speech types differed in durations and pausing, where S used values predicted from the text, while M was synthesised using durations and pauses from the natural speech (N), as an attempt to create synthetic speech with more lifelike prosody. There are six ways to assign the three speech types to the three interviews in a one-to-one manner. There are also six different orders in which the interviews can be presented, thus requiring $6 \times 6 = 36$ listeners for the fully balanced design we used.

1. How many people have tried the cross-channel swim? • *7000* • 7500 • 9000 • 6000
2. What did David not mention buying in charity shops? • suits • *coats* • shirts • ties

Table 1: Example questions from the David Walliams interview. The correct answer is highlighted using *italics*.

Directly after listening to an interview, the listener answered the 20 corresponding multiple-choice questions. The order of the questions and corresponding response options were both randomised. Since questions were asked after the entire 10-minute interview, the effects of listener memory, retention and internalisation also factored into the measured comprehension. However, because the interviews are approximately the same length and were presented in a balanced design, we expect individual variation in such traits to cancel out.

The 36 listeners were seated in sound isolated booths and listened to the interview excerpts using Beyerdynamic DT 770 PRO headphones. The questions were presented on a computer screen, where listeners selected their answer from a drop-down list. Listeners were remunerated for their time and effort.

After completing the experiment, listeners filled in a short questionnaire. The questionnaire asked how familiar the listener was with Desert Island Discs and with each of the four speakers. They were also asked for their opinion on how difficult it was to answer the questions and given the opportunity to share any other comments and observations they may have had.

3.3. Creation of synthetic stimuli

3.3.1. Model training

Three speech synthesisers were trained following established recipes using standard purpose-recorded speech databases. A single-speaker Scottish-accented female database was used to train the voice of the interviewer, and single-speaker British male and female databases were used for the voices of the guests. Note that the same TTS (text-to-speech) model was used for both male guests. The databases consisted of 238, 64 and 96 minutes of speech, respectively, excluding silent portions at the beginning and ends of sentences, and were sampled at 48 kHz.

Training of both synthesis front-ends and acoustic models broadly followed the description given in [25]. Accent-specific variants of the Combilex lexicon [26] were chosen to match each of the TTS voice talents' accents. The speech data was automatically aligned with text-derived annotation using forced alignment with 5-state hidden Markov models allowing for the insertion of pauses between words.

Deep neural networks (DNNs) were trained to predict both the duration of phones and frames of acoustics. The duration models were trained to map from phone-level inputs ex-

	N	S	M
DW	9:56	9:25	9:28
SC	10:00	11:13	10:13
VW	10:08	12:25	10:02

Table 2: Durations (in minutes and seconds) of natural and synthetic interview excerpts used.

tracted from the annotation provided by the front-end to 5-dimensional vectors indicating the frame durations of the states in a phone. Input features included binary features encoding the phone identities and phonetic features of phones in a 5-phone window centred around the phone for which predictions were to be made, and the stress and syntactic categories of the syllables and words in 3-unit windows centred around the target unit. Continuous-valued features recorded size and positional information (such as the number of syllables till the end of the word, or the number of words in the current phrase). The acoustic models were trained to map from frame-level inputs; for these, the linguistic features used for the duration model were supplemented with the index of the current sub-phone state (obtained from forced alignment) and the fraction of frames passed since the start of the current state. The outputs of the acoustic model were features extracted using STRAIGHT [27]: 60-dimensional mel-cepstral coefficients, 25 band aperiodicities and logarithmic fundamental frequency ($\log F_0$).

3.3.2. Speech generation

Stimuli for conditions S and M were synthesised from a manually-checked transcript of the data. Some care was taken to ensure that false starts and filled pauses were recorded in the transcripts. The audio was segmented into sentences, each assigned to the guest or the host speaker. The transcripts were passed through the TTS front-ends, and the resulting annotation was used in two different ways, depending on the condition: For the completely synthetic condition S, the front-end's predictions of sentence-internal pauses were used directly, durations and then acoustic features were predicted with the two DNNs for each voice. For the duration-modified condition M, the forced alignment models used to obtain aligned training data were used to align the test-set annotation with the interview audio. Note that for this reason, the alignment model used 12-dimensional MFCCs plus energy with dynamic features appended, and utterance-level cepstral mean normalisation. These are different from the synthesis features, but extracted at a frame rate compatible with them (5 ms), and are sufficiently speaker-independent to produce an alignment which performed well on the speakers in the Desert Island Disc data, as long speech was free of background noise. Ideally, the annotation created in this way would allow the generation of speech which is synthetic in all respects, except for segmental durations and placement of sentence-internal pauses.

After sentences were generated by the appropriate model for each speaker, they were concatenated into whole-interview extracts for use in the listening test. Inter-sentence pauses having the same duration as the natural ones but consisting of pure silence were inserted between sentences during the concatenation. Table 2 lists the durations of the resulting interview excerpts under the different conditions.

4. Results

This section presents our experimental results and analysis; deeper discussion and interpretation is reserved for Section 5.

	N	S	M	All types
DW	164/240 68%	181/240 75%	134/240 56%	479/720 67%
SC	176/240 73%	144/240 60%	147/240 61%	467/720 65%
VW	190/240 79%	181/240 75%	157/240 65%	528/720 73%
All int.	530/720 74%	506/720 70%	438/720 61%	1474/2160 68%

Table 3: Number and fraction of correct responses across speech types (columns) and interviews (rows).

Comparison	**N vs. S**	**N vs. M**	**S vs. M**
Difference	3.3%	13%	9.4%
Adjusted p-value	0.18	$< 10^{-6}$	$4.0 \cdot 10^{-4}$

Table 4: Differences in the overall rate of correct response between speech types, and adjusted p-values for a null hypothesis that the difference is zero.

The results of the listening test are presented in Table 3. A total of 240 responses (12 listeners $\times$ 20 questions) were collected for each interview in each speech-type condition. The rate of correct response is seen to vary substantially across interviews and speech types. To get a clearer picture of the effect of speech type on the average rate of correct response, one can pool the different interviews as seen in the last row of Table 3. The experiment was carefully balanced so that this pooling does not introduce bias. Table 4 reports on differences in the total rates of correct response between different speech types, to quantify the size of any comprehension effect in the experiment. The table also lists p-values from two-tailed Fisher's exact tests to assess the degree of significance of the observed differences, with the Holm-Bonferroni method [28] applied to adjust the significances for multiple comparisons.

In the post-test questionnaire, 13 listeners (group 'D', for 'difference') reported that the multiple-choice questions were easier to answer for natural speech than for synthetic speech, while 23 listeners (group 'ND') did not report a difference in perceived difficulty. Of the 13 listeners in group D, 8 listened to natural DW (and, thus, synthetic SC and VW), 2 to natural SC, and 3 to natural VW. There were no reports of synthetic speech questions being easier than natural speech.

The subjective, reported difficulty in answering questions for different speech types can be compared against the objective performance on the multiple-choice questions. Figure 1 shows a box plot of the difference in rate of correct response on the natural speech (N) minus the synthetic speech (S and M pooled) for the listener groups ND and D. Spearman's rho, a kind of non-parametric correlation coefficient for ordinal variables, shows a weak but positive association of $\rho = 0.23$ ($p = 0.17$) between the objective performance difference and the reported difference in question-answering difficulty (ND or D).

Figure 2 shows a scatterplot of listener's performance of natural (N) and synthetic (S and M) speech, thus illustrating both the range and distribution of listener's correctness scores in natural and synthetic speech types, and their relative difference. (Listeners who performed better on natural speech are above the dotted line.) Listeners in the groups ND or D are distinguished using different symbols.

Table 5 summarises how familiar listeners reported being with the different speakers from the underlying interview material (the four possible responses have been converted to ordinal

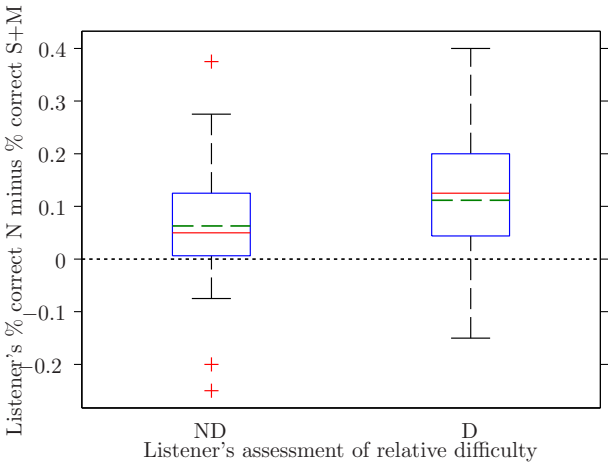

Figure 1: Performance differences between natural and synthetic speech for listeners in groups ND and D. Red lines are medians, green dashed lines are means; box edges are at 25 and 75% quantiles.

Familiarity	0	1	2	3
KY (host)	18	8	8	2
DW	2	5	17	12
SC	7	13	9	7
VW	17	12	4	3

Table 5: Raw counts of listeners' reported familiarity with the speakers appearing in the interviews. 0 is least familiar ("don't know who s/he is") while 3 is the most familiar ("know her/him really well"). All rows sum to 36.

labels for compactness). To assess whether speaker familiarity may have affected the rate of correct response, the set of each listener's rate of correct response on each synthesised (S or M) interview were grouped according to the reported familiarity with the interviewee whose voice had been replaced with a synthetic speaker. Spearman's rho between this reported familiarity and the objective correctness score is $\rho = -0.006$ ($p = 0.96$), indicating no relation.

5. Discussion

The goal of the evaluation was to measure comprehension differences between synthetic and natural speech using a post-perceptual approach. Our results paint a complex picture. Overall, it is clear that subjects perform significantly worse on modified synthesis (M) than on regular synthetic or natural speech, even though one might have expected it to fall between S and N due to the natural durations used. We suspect this may be due to mismatch between the training and test data. In particular, acoustic models learned on the carefully paced read-speech training material may not produce highly intelligible or comprehensible speech when shoehorned into the spurt-like duration structure of the interview speech. In addition, overlapping speech and laughter tended to have a very detrimental effect on the automatic alignment, which may have played a role as well.

While average comprehension performance on regular synthetic speech (S) is 3.3% lower than on natural speech N, this difference is not statistically significant. These results suggest that post-perceptual tests at present are not sensitive enough to easily identify comprehensibility differences, even when using prosodically rich conversational material.

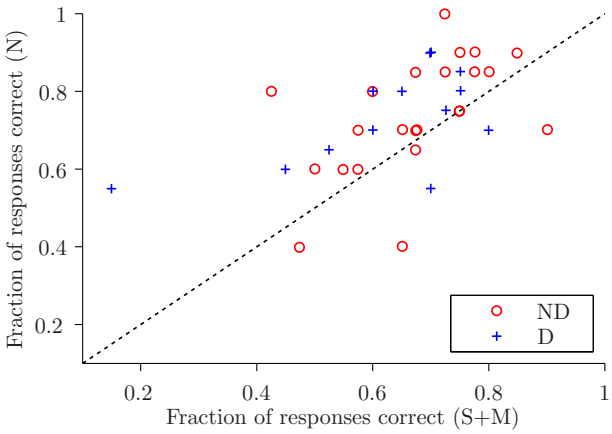

Figure 2: Scatterplot of listener accuracies on natural and synthetic speech. Points have been jittered slightly along the axis $x = y$ (dotted) to separate overlapping symbols.

The detailed performance breakdown in Table 3 presents a complex pattern of performance differences across interviews and speech types. The numbers suggesting that synthesised (S) SC is much more challenging than synthesised DW, compared to their respective natural (N) interviews, may appear difficult to reconcile with the fact that the synthetic voice used for both these interviewees is exactly the same. However, while large in magnitude, most of the differences between N and S in the table are not significant due to the small sample sizes prior to pooling. This and other indications suggest that much of the observed complexity may simply be attributable to random variation.

In their comments, many participants said that synthetic speech was more difficult to focus on, but also that despite this they found the task do-able, which is supported by the objective data. Two participants even described being 'nauseated' by listening to the synthetic speech. Taken together, this suggests that the experience of listening to extended segments of synthesised speech is quite degraded compared to natural speech, though not in a manner that stands out in our objective test.

In future work, we will investigate our question sets in more detail. The segmental intelligibility of the three types of speech will be measured by giving listeners a multiple-choice question and then playing the relevant fragment of speech containing the answer to the question. This will indicate whether M was simply less intelligible than S, or if other factors are at play as well.

Developing ecologically-relevant evaluation techniques for synthetic speech is still very much a work in progress. The type of scenarios we increasingly need novel techniques for are for example: "How to evaluate a voice built on found data like Desert Island Discs?" or "How should audio books or conversational voices be evaluated?" Recent years have seen a few attempts to develop evaluation techniques for these scenarios, e.g., evaluation of the personality of synthetic voices [29], psycholinguistic studies into filled pauses for synthetic speech [30] and the evaluation of intonation [31]. However, there is still a lot of room to investigate and design more appropriate evaluation strategies.

Acknowledgements This work was supported by the EPSRC under Programme Grant EP/I031022/1 (Natural Speech Technology). The NST research data collection may be accessed at http://datashare.is.ed.ac.uk/handle/10283/786. The data for this paper will be made publicly available upon paper acceptance.

6. References

[1] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *Blizzard Challenge Workshop 2012*.

[2] S. King, "Measuring a decade of progress in Text-to-Speech," *Loquens*, vol. 1, no. 1, p. e006, 2014.

[3] A. Cutler, D. Dahan, and W. Van Donselaar, "Prosody in the comprehension of spoken language: A literature review," *Language and speech*, vol. 40, no. 2, pp. 141–201, 1997.

[4] L. M. Slowiaczek and H. C. Nusbaum, "Effects of speech rate and pitch contour on the perception of synthetic speech," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 27, no. 6, pp. 701–712, 1985.

[5] C. R. Paris, M. H. Thomas, R. D. Gilson, and J. P. Kincaid, "Linguistic cues and memory for synthetic and natural speech," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 42, no. 3, pp. 421–431, 2000.

[6] M. Marković, B. Jakovljević, T. Milićev, and N. Milievć, "The role of prosody in the perception of synthesized and natural speech," in *Speech and Computer*. Springer, 2015, pp. 446–453.

[7] M. Charfuelan and I. Steiner, "Expressive speech synthesis in MARY TTS using audiobook data and EmotionML," in *Interspeech*, 2013, pp. 1564–1568.

[8] L. Chen, M. J. F. Gales, V. Wan, J. Latorre, and M. Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training." in *Interspeech*, 2012.

[9] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, and S. King, "TUNDRA: a multilingual corpus of found data for TTS research created with light supervision." in *Interspeech*, 2013, pp. 2331–2335.

[10] A. Gallardo-Antolín, J. M. Montero, and S. King, "A comparison of open-source segmentation architectures for dealing with imperfect data from the media in speech synthesis," in *Interspeech*, 2014.

[11] D. B. Pisoni, L. M. Manous, and M. J. Dedina, "Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility," *Computer Speech & Language*, vol. 2, no. 3, pp. 303–320, 1987.

[12] M. E. Reynolds, C. Isaacs-Duvall, and M. L. Haddox, "A comparison of learning curves in natural and synthesized speech comprehension," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 4, pp. 802–810, 2002.

[13] J. J. Jenkins and L. D. Franklin, "Recall of passages of synthetic speech," *Bulletin of the Psychonomic Society*, vol. 20, no. 4, pp. 203–206, 1982.

[14] J. V. Ralston, D. B. Pisoni, S. E. Lively, B. G. Greene, and J. W. Mullennix, "Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 33, no. 4, pp. 471–491, 1991.

[15] D. B. Pisoni and S. Hunnicutt, "Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system," in *ICASSP*, vol. 5, 1980, pp. 572–575.

[16] T. Boogaart and K. Silverman, "Evaluating the overall comprehensibility of speech synthesizers," in *Second International Conference on Spoken Language Processing*, 1992.

[17] K. Silverman, S. Basson, and S. Levas, "Evaluating synthesiser performance: is segmental intelligibility enough?" in *First International Conference on Spoken Language Processing*, 1990.

[18] C. Delogu, S. Conte, and C. Sementina, "Cognitive factors in the evaluation of synthetic speech," *Speech Communication*, vol. 24, no. 2, pp. 153–168, 1998.

[19] Y.-Y. Chang, "Evaluation of TTS systems in intelligibility and comprehension tasks," in *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing*, 2011, pp. 64–78.

[20] D. J. Higginbotham, A. Drazek, K. Kowarsky, C. Scally, and E. Segal, "Discourse comprehension of synthetic speech delivered at normal and slow presentation rates," *Augmentative and Alternative Communication*, vol. 10, no. 3, pp. 191–202, 1994.

[21] G. P. Sonntag, T. Portele, and F. Haas, "Comparing the comprehensibility of different synthetic voices in a dual task experiment," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[22] S. A. Duffy and D. B. Pisoni, "Comprehension of synthetic speech produced by rule: A review and theoretical interpretation," *Language and Speech*, vol. 35, no. 4, pp. 351–389, 1992.

[23] S. J. Winters and D. B. Pisoni, "Perception and comprehension of synthetic speech," *Research on spoken language processing report*, no. 26, pp. 95–138, 2004.

[24] BBC. "Desert Island Discs Podcasts". [Online]. Available: http://www.bbc.co.uk/programmes/b006qnmr

[25] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Interspeech*, 2015.

[26] K. Richmond, R. Clark, and S. Fitt, "On generating Combilex pronunciations via morphological analysis," in *Interspeech*, 2010, pp. 1974–1977.

[27] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[28] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[29] M. Wester, M. Aylett, M. Tomalin, and R. Dall, "Artificial personality and disfluency," in *Interspeech*, 2015.

[30] R. Dall, M. Wester, and M. Corley, "The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech," in *Interspeech*, 2014.

[31] J. Latorre, K. Yanagisawa, V. Wan, B. Kolluru, and M. J. F. Gales, "Speech intonation for TTS: Study on evaluation methodology," in *Interspeech*, 2014.